The ICSI 2007 Speaker Diarization System

Chuck Wooters[†] and Marijn Huijbregts[†,‡]

[†]International Computer Science Institute
Berkeley, California, USA

[‡]Department of Electrical Engineering, Mathematics and Computer Science,
University of Twente, The Netherlands

Outline

- Submitted Tasks
- Eval results
- System description
- What's new/different from RT-06s
- Post-evaluation analysis
- Future work

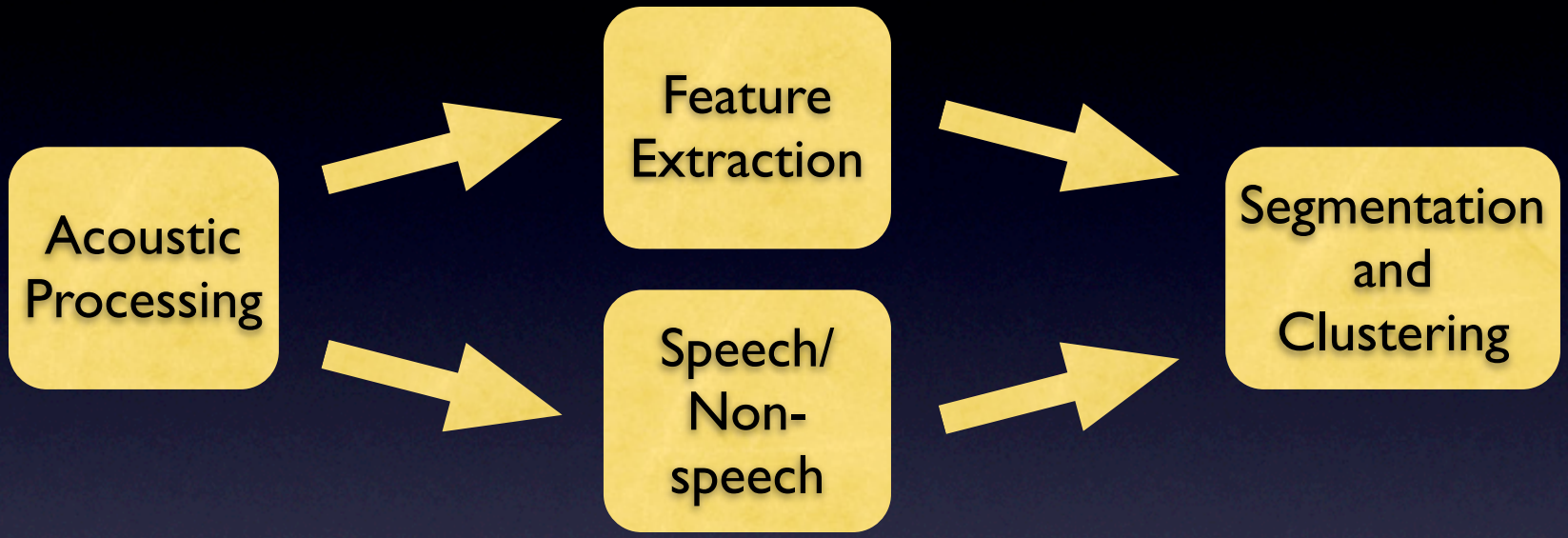
Tasks

- Conference room
  - MDM
  - SDM
- No lecture or coffee-break

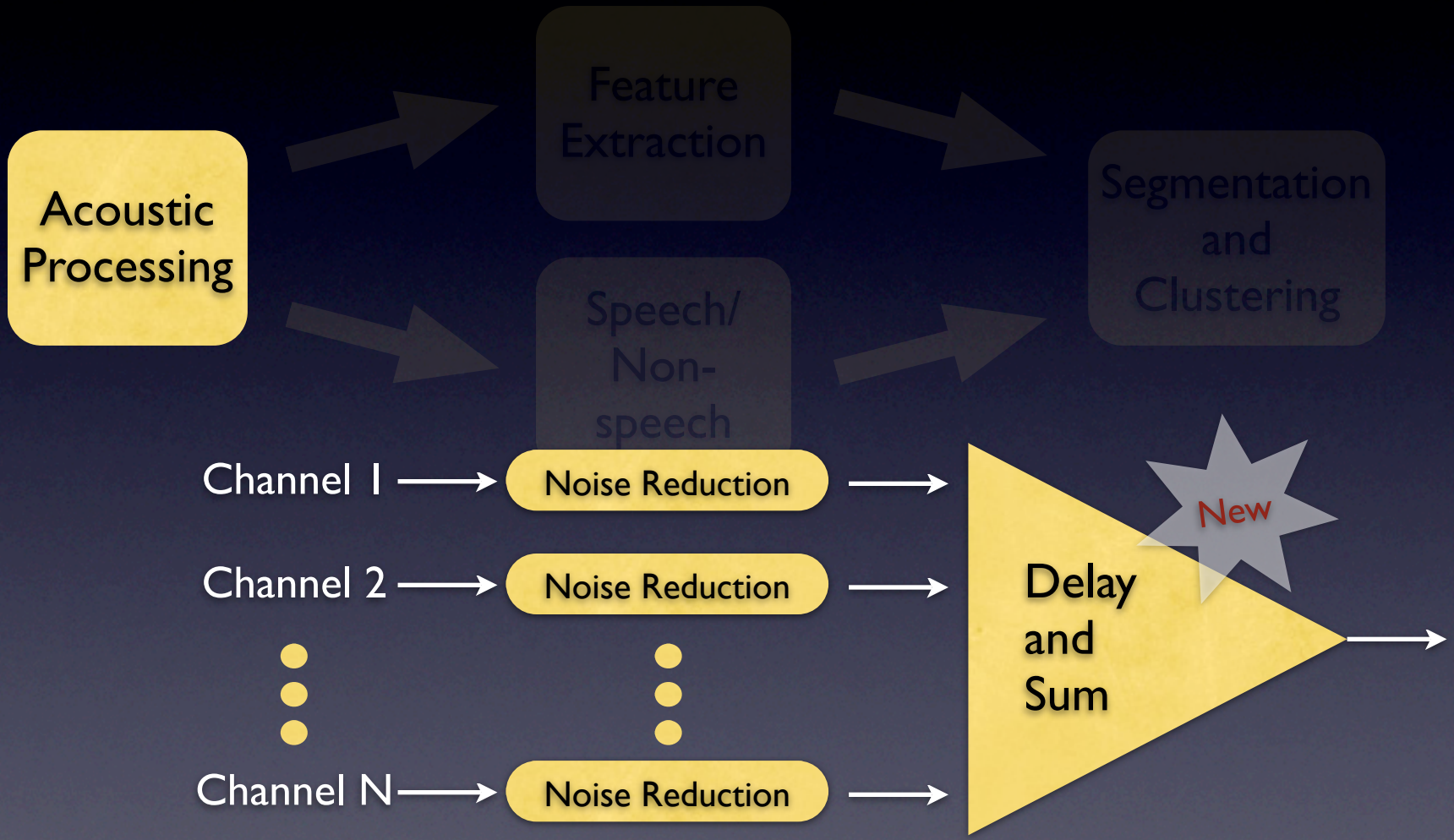
Official Eval07 Results

	%Miss	%FA	%Spkr	%DER
MDM	4.5	1.5	2.5	8.51
SDM	5.0	1.8	14.9	21.74

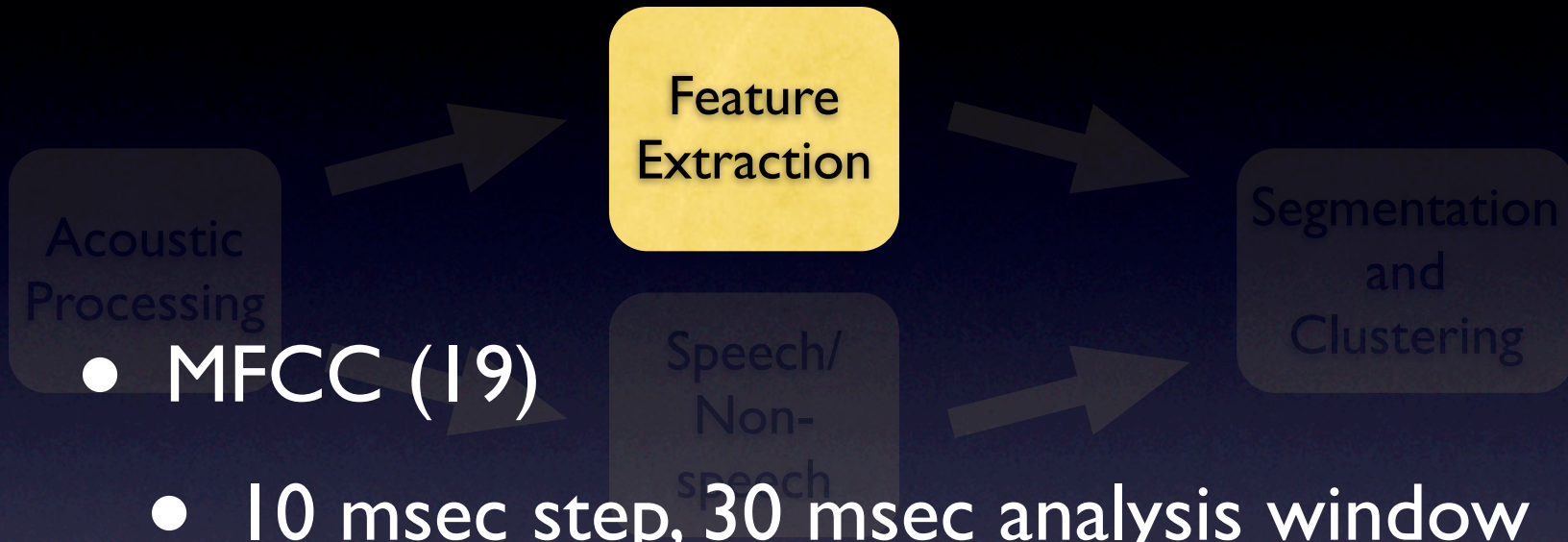
System Description



System Description

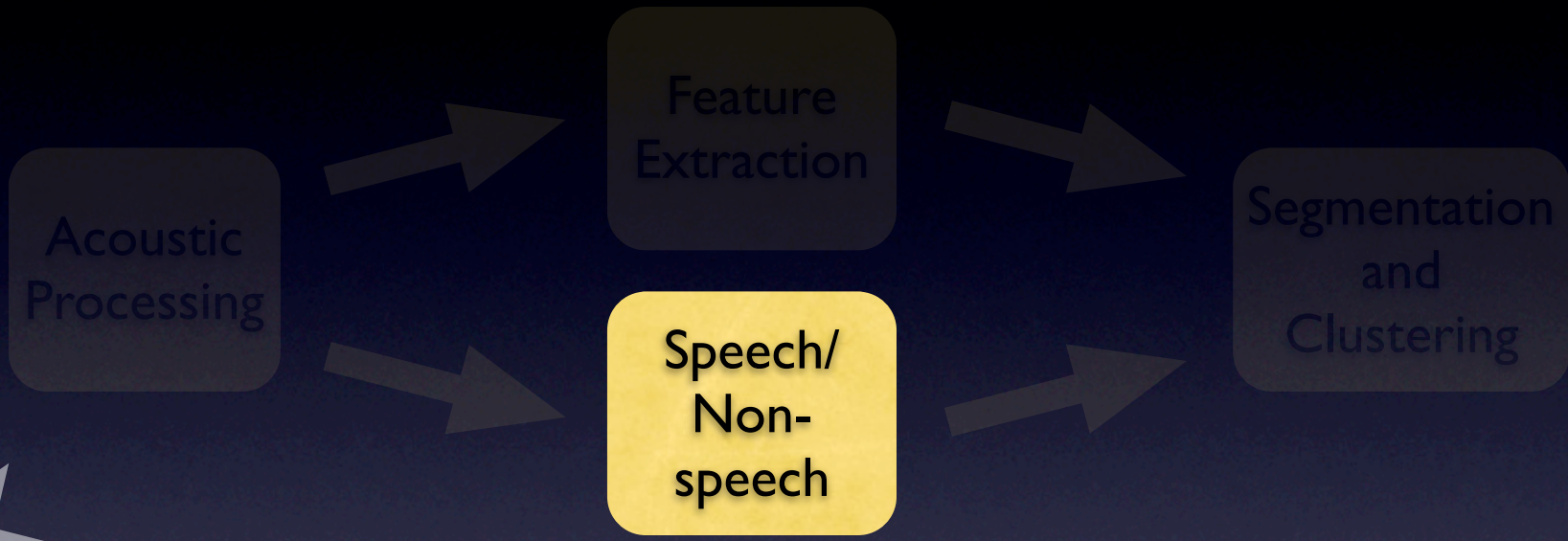


System Description



- MFCC (19)
  - 10 msec step, 30 msec analysis window
  - Use HTK (HCopy) to generate features
- For MDM
  - Run delay and sum again with 10 msec step to generate delay features

System Description



New

1. Initial "guess" at speech and non-speech regions.
2. Iterative re-segmentation and training into "silence", "sound", and "speech" models.
   - Low-energy non-speech regions: "silence"
   - Remaining non-speech regions: "sound"
   - Speech regions: "speech"
3. Use BIC to detect "sound" == "speech", if so, merge

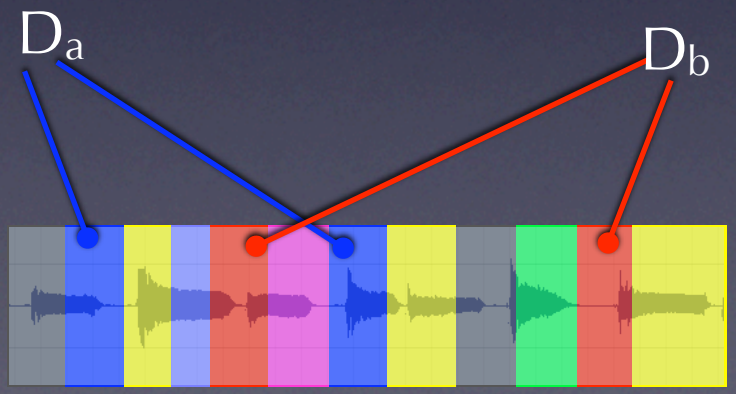
System Description

- Agglomerative Clustering
- 16 initial clusters
  - linear initialization
- Each cluster modeled using a diagonal covariance GMM
- 5 initial Gaussians per cluster
- 2.5 sec min segment duration
  - Final alignment pass with 1.5 sec min

System Description

- Modified BIC for deciding which clusters to merge and when to stop clustering
- No explicit penalty term

Segmentation and Clustering



$$\log p(D|\theta) \geq \log p(D_a|\theta_a) + \log p(D_b|\theta_b)$$

with

$$D = D_a + D_b \quad \text{and} \quad M_\theta = M_{\theta_a} + M_{\theta_b}$$

What else is new this year

- Larger development set
- New stream-weighting algorithm
- Smoothing after clustering, not before

Larger Dev Set

- Our hand-picked "dev06" data (12 meetings)
- Added eval06 data (9 meetings)
- Large dev set important for "smoothing" diarization results
- Results on larger dev set led to removal of:
  - Frame "pruification"
  - "Friends and Enemies"
  - Automatic estimation of number of clusters

New Stream-weighting

- Previously used fixed stream weights
- Weights now determined dynamically based on the variance in the BIC scores for each stream (see X. Anguera's thesis)
- Initial "guess" at stream weights required (0.65 for mfcc stream)

	Eval07 %DER
Dynamic	8.51
Fixed (0.9/0.1)	9.29

Post-clustering smoothing

- Gaps of < 0.5 seconds between same-speaker segments are removed using NIST's rttmSmooth-v3.pl script.
- Helps a little by reducing missed speech (with a small increase in false alarms)

Smoothing	%Miss	%FA	%Spkr	%DER
0.5	4.5	1.5	2.5	8.51
0.3	5.2	1.1	2.5	8.82

Note: no longer smoothing *before* clustering.

Post-eval Analysis

Speech/Non-Speech and Overlap Effects

MDM

- Score with (+) and without (-) overlap, and with and without ref sp/nonsp input

	%Miss	%FA	%Spkr	%DER
-ref,+ovlp (base)	4.5	1.5	2.5	8.51
+ref, +ovlp	3.7	0.0	3.8	7.47
-ref, -ovlp	0.9	1.6	2.6	5.11
+ref, -ovlp	0.0	0.0	3.9	3.94

Overlap = ~3.5%, Sp/Nonsp = ~2.5%, Spkr = ~2.5%

SDM

- Score with (+) and without (-) overlap, and with and without ref sp/nonsp input

	%Miss	%FA	%Spkr	%DER
-ref,+ovlp (base)	5.0	1.8	14.9	21.74
+ref, +ovlp	3.7	0.0	12.8	16.51
-ref, -ovlp	1.4	2.0	14.7	18.03
+ref, -ovlp	0.0	0.0	12.7	12.75

Overlap = ~3.5%, Sp/Nonsp = ~3.5%, Spkr = ~15%

Delay Features

- What is the gain from using delay features as a second stream?

	Eval07 %DER
MFCC only	14.02
MFCC+Delays	8.51

Noise Filtering

- What is the gain from the noise reduction (Wiener filtering)?

	Eval07 %DER	Eval07 SAD Error
No Filtering	15.80	3.4
Filtering	8.51	3.3

Noise Filtering II

- What if we apply filtering sometimes?

Where do we filter?	Eval07 %DER
-None-	15.80
SAD	10.54
SAD & MFCC	12.99
SAD & Delays	13.70
SAD & MFCC & Delays	8.51

Future Work

- Work on overlaps (especially for MDM)
- Additional feature streams (especially for SDM)
- Work on speech/non-speech- where are the errors coming from?
- Significance testing for diarization

Significance Testing for Diarization

- Based on NIST's Matched Pairs Sentence-Segment Word Error (MAPSSWE) test
- Find optimal mapping between system RTTMs and reference RTTM
- Create "words" by sampling from sys and ref RTTMs at a desired interval
- Compute the Z-score on the differences, as in MAPSSWE
- Initial tests are consistent with intuitions...